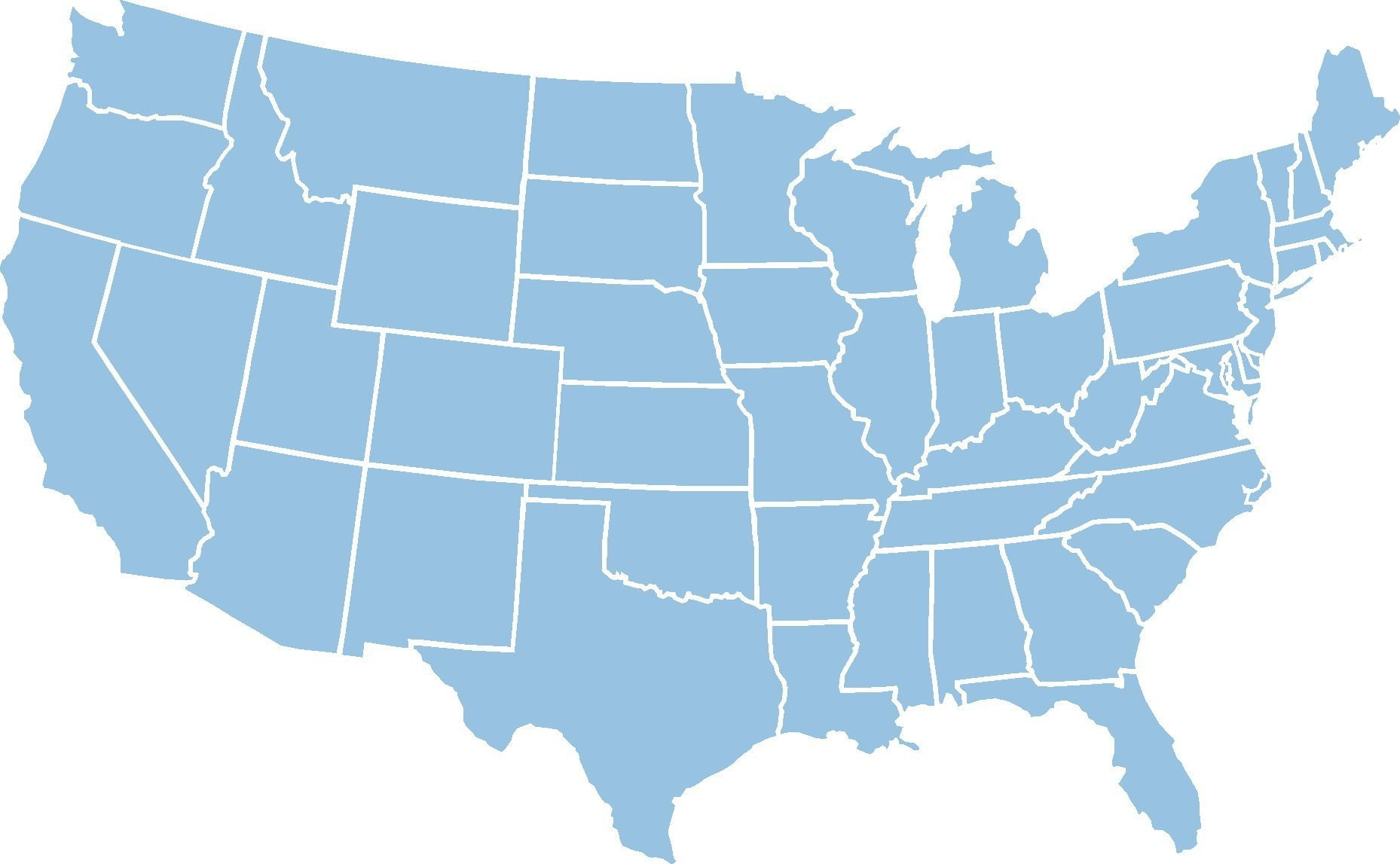
# Communities & Crime:

A Regression Analysis



Ryan Flynn

Theresa Van

Stephen Shu

Eric Muro

California State Polytechnic Univ., Pomona

# 

# Introduction:

In the abstract for the data set, the objective was clearly stated that we were to use this set to predict violent crimes per capita in a given community and were presented with predictors “if there was any plausible connection to crime (N=122), plus the attribute to be predicted (Per Capita Violent Crimes)”. The data set was initially tailored for the WEKA machine learning software, but was reformatted to be suitable for analysis in R. After manipulation, we were able to obtain data sets containing 102 predictors (p-removed) and 125 (n-removed) variables respectively (see “Missing Data”).

# Data Description:

### Data Source(s):

The data sources from two major sources: community driven data is built from US census data (updated every ten years) while data relevant to crime pulls from a Law Enforcement Management and Statistics (LEMAS) survey from the Bureau of Justice Statistics (BJS). It should be noted that some data related to crime only captures information for communities employing over 100 active duty police officers, so extrapolation to small-town communities or communities with low officer employment is not possible using this model.

### Missing Data:

For this data set, there were issues with a practical analysis due to the presence of missing data. To remedy this, we manipulated the data set by creating two separate, truncated data sets. One set excluded predictors (i.e. columns) which contained empty or missing values initially represented by a ‘?’, while the other performed the same procedure row-wise. Following this change to the data, we could perform a more robust analysis to predict violent crime per capita.

#### Preliminary Analysis:

Initially, a scatterplot matrix was chosen to identify any issues with collinearity of predictors. However the large pool of predictors made such a process unexecutable in R. We used the VIF factor to detect multicollinearity and to detect the predictor with the highest VIF value, remove it, recalculate the VIF values for all remaining predictors, and repeat until all the remaining predictors have a VIF value below 10. Following the removal of high-VIF terms, a full linear model was regressed on using 59 predictors for our response, splitting data into training and test by randomly sampling data into each set by using 50 percent of the full set.

|  |  |
| --- | --- |
| Model | Adjusted R Squared |
| Full LM | 0.6621 |
| Full LM w/ State Interaction | 0.6699 |

Figure 1: Selection Criteria for Full Model with Interaction Term

Using adjusted R Squared as our basis criteria in our preliminary analysis of interaction terms, it was observed that there was not a significant increase in the criterion. In favor of interpretability, this model was not selected initially. Moreover, since the residual plots were so similar to Figure 1, interpretability again prevailed as the preferred selection criterion. Further analysis below shows that neither of these models were good enough at predicting on test data.

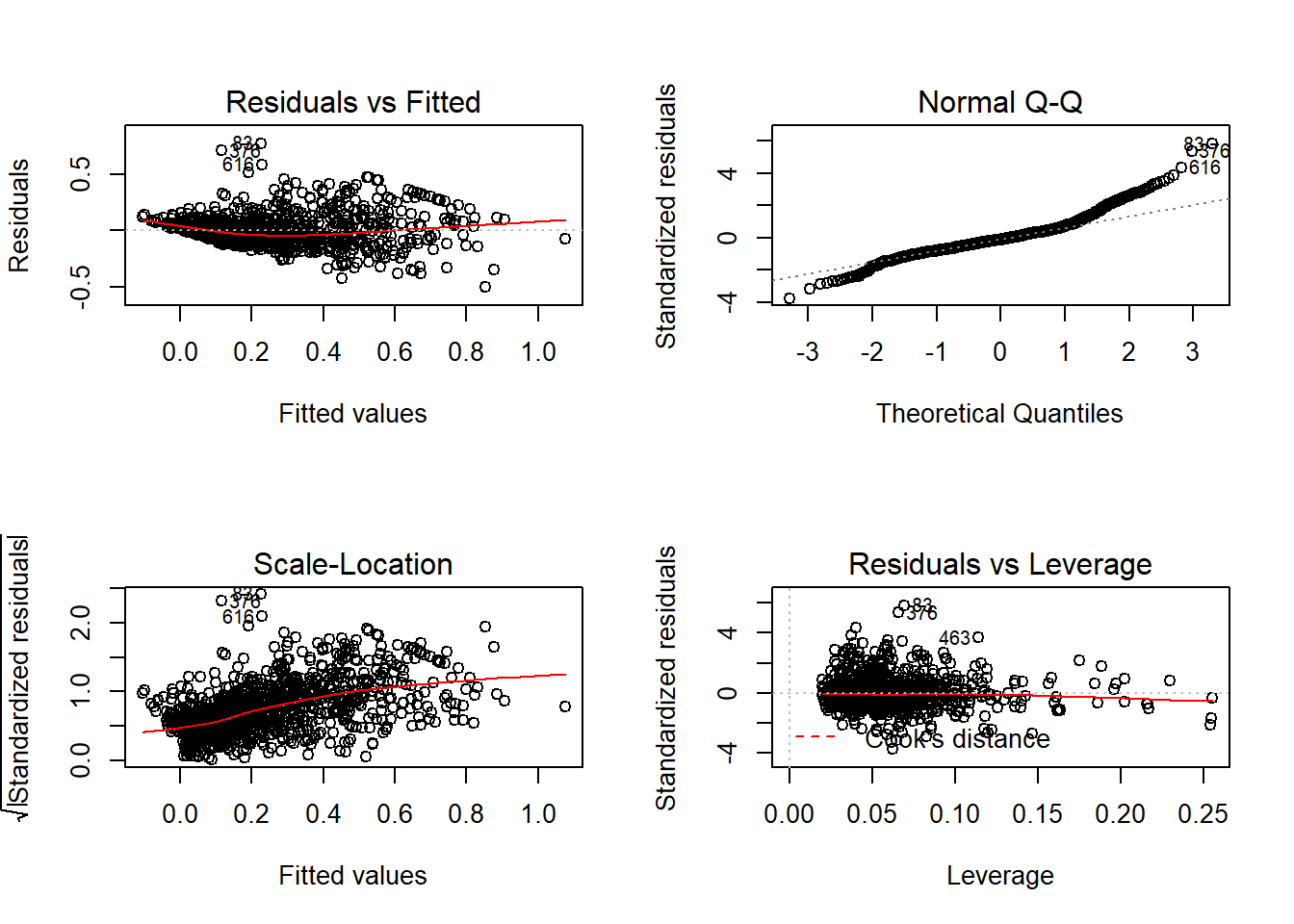


Figure 2: Summary Residual Plots of Full Linear Model

Above are residual plots for the full model with 59 predictors. There is a small clustering of outliers as indicated by the clustering of points in the Residuals v. Fitted plot (top left) above the zero line. There are several leverage points at the far right of the Residuals vs Leverage plot (bottom right).

For the reasons above, and because the full linear model presented few significant predictors, it was not chosen as an ideal model, but provided a good starting point for deeper analysis. As seen below, lasso regression eventually became a more optimal option as many predictors were not ultimately related to the response.

#### Model(s) Selection:

We consider forward stepwise subset selection models, lasso regression models, and ridge regression models. Figure 3 below represents a criteria selection matrix highlighting (in yellow) minimal values across both data sets (red: n-removed and blue: p-removed) using Cp, BIC, and Adjusted R Squared as basis criteria, in addition to test and validation set error.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Type | p | Cp | BIC | Adjusted R Sq. | LOOCV | Validation Set MSE |
| Tm1 | BS | 26 | 19.6565 | -946.0668 | 0.6645755 | 0.01895667 | 0.01898845 |
| Tm2 | BS | 12 | 43.54231 | -990.6249 | 0.6517371 | 0.01927082 | 0.01894992 |
| Tm3 | BS | 36 | 26.71804 | -890.3419 | 0.6656187 | 0.01898466 | 0.0190881 |
| Tm4 | Ridge | 58 | --------- | --------- | --------- | 0.01863495 | 0.01894046 |
| Tm5 | Lasso | 42 | --------- | --------- | --------- | 0.01801492 | 0.01745593 |
| Em1 | BS | 8 | -12.93019 | -136.3232 | 0.6292293 | 0.02946906 | 0.02798789 |
| Em2 | BS | 21 | -3.014478 | -123.6918 | 0.6371149 | 0.02887488 | 0.02750994 |
| Em3 | Ridge | 61 | --------- | --------- | --------- | 0.03096675 | 0.03066004 |
| Em4 | Lasso | 23 | --------- | --------- | --------- | 0.02491547 | 0.02217443 |

Figure 3: Selection Criteria Matrix for BS, Lasso, and Ridge Models from n-removed and p-removed data sets

It should be noted that obtaining the residual plots for the Lasso and Ridge regression models were outside the scope of the **glmnet** package that was used for building our models and were thus excluded from this report. In spite of this, we see it produces the best results for test and validation set error on the n-removed data set. Across the board, p-removed methodology produced a series of models which were each minimal across criterion in each of the different models.

Also outside of the capabilities of these models is a comparison basis between the two different data sets because they pull from different training data sets. A means by which to obtain a training data set consistent with both sets of data was not achievable in the scope of this project, since n-removed data contained predictors which p-removed data did not contain.

# Final Prediction Results:

# When deciding to choose between n-removed and p-removed, preference was given to n-removed as it produced models with consistently low criteria across the board. Moreover, removing predictors also creates a potential problem of reducing accurate model prediction on test set, especially if significant predictors are removed merely on the basis of containing missing values. From the initial full linear model, it was determined the final model selection was made relatively simple using Lasso method (Em4), as many predictors ultimately were not related to the response. Intuition is supported here by data, as Figure 3 shows a minimal test and validation set error relative to the other p-removed models.

#### Coefficients:

As seen below in Figure 4 some coefficients with high association with violent crime per capita were Percentage of Divorced, or Never Married, males, number of homeless population living on the street, the percentage of houses with less than 3 bedrooms, percentage of unemployed persons and percentage of houses without a home phone.

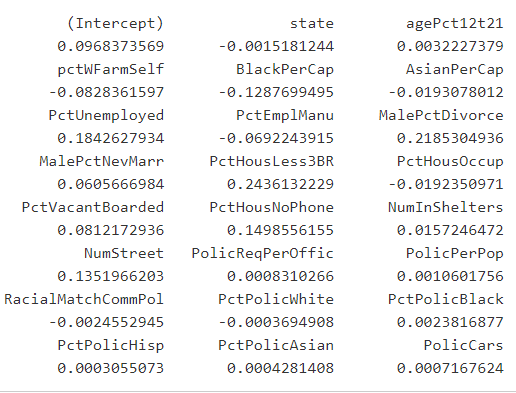


Figure 4: Non-zero lasso regression coefficients

# Conclusion:

The biggest challenge in prediction modeling, as with any sort of data analysis, was the issue of missing data. The decision to exclude incomplete data points was one which dramatically affects the outcome of the models produced. In spite of this, modeling was still possible and produced some coefficients which appear to be match common intuition (i.e. unemployment, homelessness, divorce status). In terms of the response, one of the interesting aspects looking forward with this model is the changes which could come with the influx of new census data that will be coming in the next few years.

Moving forward, a deeper analysis into finding methods to compare both series of models using a consistent training set would be ideal.